Network inference for truncated gaussian data

Anne Gégout-Petit, Aurélie Gueudin, Clémence Karmann

Network inference for truncated gaussian data

Anne Gégout, Aurélie Gueudin and Clémence Karmann

Université de Lorraine, Inria Nancy Grand Est, FRANCE

anne.gegout-petit@univ-lorraine.fr

EMS 2019
Palermo

Introduction

- Infer network defined by conditional dependencies

Introduction

- Infer network defined by conditional dependencies

- Gaussian variables zero inflated by double truncation

Introduction

- Infer network defined by conditional dependencies

- Gaussian variables zero inflated by double truncation
- Applications for the modelling of interactions between microbacterial populations
  - left truncation: replication phenomena
  - right truncation: non restrictive hypothesis for theoretical results !

Content

1. Modelling
2. Estimation process
   - Covariance matrix estimation
   - Precision matrix estimation
3. Theoritical results
4. Simulation studies

Modelling

Modelling

Let $\mathbf{X} \sim \mathcal{N}_p(0, \Sigma^*)$ be a gaussian p-vector with $\Sigma^*_{jj} = 1$ for all j. $\mathbf{X}$ is not observed but we observe $\mathbf{Y}$ defined from $\mathbf{X}$ by:
$Y_j = \mathbb{1}_{a_j \leq X_j \leq b_j} X_j$ where $a_j < b_j$.

A non directed graph structure is given by the inverse of the covariance matrix $\Theta^* := (\Sigma^*)^{-1}$:

$$\begin{aligned} X_j \longleftrightarrow X_k &\iff X_j \not\perp\!\!\!\perp X_k \mid (X_l)_{l \neq j,k} \\ &\iff \mathrm{cor}(X_j, X_k \mid (X_l)_{l \neq j,k}) \neq 0 \\ &\iff \Theta^*_{jk} \neq 0. \end{aligned}$$

Aim

Aim

Infer the latent graph structure given by the precision matrix Θ^* of the gaussien vector $\mathbf{X}$ from the observations of the truncated vector $\mathbf{Y}$.

Theorotical tools - Pairs likelihood

Let us look at the likelihood $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ of the pairs $(Y_j, Y_k),\ j < k$ where :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k),$$

with:

Theorotical tools - Pairs likelihood

Let us look at the likelihood $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ of the pairs (Y_j, Y_k), $j < k$ where :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k),$$

with:

- $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j = a, \zeta_k = b}$ où $\zeta_l = \begin{cases} 1 \text{ si } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ si } y_l = 0. \end{cases}$

Theorotical tools - Pairs likelihood

Let us look at the likelihood $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ of the pairs $(Y_j, Y_k),\ j < k$ where :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k),$$

with:

- $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j=a,\zeta_k=b}$ où $\zeta_l = \left\{ \begin{array}{l} 1 \text{ si } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ si } y_l = 0. \end{array} \right.$
- $\phi_{11,jk}(\sigma, y_j, y_k) = f(y_j, y_k, \sigma)$

Theorotical tools - Pairs likelihood

Let us look at the likelihood $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ of the pairs (Y_j, Y_k), $j < k$ where :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k),$$

with:

- $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j = a, \zeta_k = b}$ où $\zeta_l = \begin{cases} 1 \text{ si } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ si } y_l = 0. \end{cases}$
- $\phi_{11,jk}(\sigma, y_j, y_k) = f(y_j, y_k, \sigma)$
- $\phi_{01,jk}(\sigma, y_j, y_k) = \phi_{01,jk}(\sigma, y_k) = \int_{[a_j, b_j]^c} f(x, y_k, \sigma)\mathrm{d}x$
- $\phi_{10,jk}(\sigma, y_j, y_k) = \phi_{10,jk}(\sigma, y_j) = \int_{[a_k, b_k]^c} f(y_j, y, \sigma)\mathrm{d}y$

Theorotical tools - Pairs likelihood

Let us look at the likelihood $\mathcal{L}_{jk}(\Sigma^*_{jk}, y)$ of the pairs $(Y_j, Y_k),\ j < k$ where :

$$\mathcal{L}_{jk}(\sigma, y) = \sum_{a,b=0}^{1} \phi_{ab,jk}(\sigma, y_j, y_k) n_{ab}(y_j, y_k),$$

with:

- $n_{ab}(y_j, y_k) = \mathbb{1}_{\zeta_j = a, \zeta_k = b}$ où $\zeta_l = \begin{cases} 1 \text{ si } y_l \in [a_l, b_l] \setminus \{0\}, \\ 0 \text{ si } y_l = 0. \end{cases}$
- $\phi_{11,jk}(\sigma, y_j, y_k) = f(y_j, y_k, \sigma)$
- $\phi_{01,jk}(\sigma, y_j, y_k) = \phi_{01,jk}(\sigma, y_k) = \int_{[a_j, b_j]^c} f(x, y_k, \sigma) \mathrm{d}x$
- $\phi_{10,jk}(\sigma, y_j, y_k) = \phi_{10,jk}(\sigma, y_j) = \int_{[a_k, b_k]^c} f(y_j, y, \sigma) \mathrm{d}y$
- $\phi_{00,jk}(\sigma, y_j, y_k) = \phi_{00,jk}(\sigma) = \iint_{[a_j, b_j]^c \times [a_k, b_k]^c} f(x, y, \sigma) \mathrm{d}x\mathrm{d}y.$

1. Modelling
2. Estimation process
   - Covariance matrix estimation
   - Precision matrix estimation
3. Theoritical results
4. Simulation studies

Context

Let $\mathbf{Y} := (Y^{(1)}, \ldots, Y^{(n)})$ be an i.i.d. n-sample du vecteur Y. For every $j < k$, let us denote:

- $\mathcal{L}_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y})$ the likelihood of the pairs sample $((Y_j^{(i)}, Y_k^{(i)}))_{1 \leq i \leq n}$ given by:

$$\mathcal{L}_{jk}^{(n)}(\sigma, \mathbf{y}) = \prod_{i=1}^{n} \mathcal{L}_{jk}(\sigma, y^{(i)}).$$

- $L_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y})$ the log-likelihood $L_{jk}^{(n)}(\Sigma_{jk}^*, \mathbf{y}) = \log(\mathcal{L}_{jk}^{(n)}(\sigma, \mathbf{y}))$

Step 1: covariance matrix Σ^*

- empirical covariance matrix from $\mathbf{Y}$: poor performance

Step 1: covariance matrix Σ^*

- empirical covariance matrix from **Y**: poor performance
- Likelihood Maximisation of **Y** : 2^p terms...

Step 1: covariance matrix Σ^*

- empirical covariance matrix from $\mathbf{Y}$: poor performance
- Likelihood Maximisation of $\mathbf{Y}$: 2^p terms...

$\rightsquigarrow$ estimator of Σ^* term to term from the pairs (Y_j, Y_k). We estimate each coefficients $\widetilde{\Sigma}_{jk}^{(n)}$ by maximising the log-likelihood of the sample $\left((Y_j^{(i)}, Y_k^{(i)})\right)_{i=1,\dots,n}$.

Step 1: covariance matrix Σ^*

- empirical covariance matrix from $\mathbf{Y}$: poor performance
- Likelihood Maximisation of $\mathbf{Y}$: 2^p terms...

$\rightsquigarrow$ estimator of Σ^* term to term from the pairs (Y_j, Y_k). We estimate each coefficients $\widetilde{\Sigma}_{jk}^{(n)}$ by maximising the log-likelihood of the sample $\left((Y_j^{(i)}, Y_k^{(i)})\right)_{i=1,\dots,n}$.

Estimator of the covariance matrix

Estimator $\widetilde{\Sigma}^{(n)}$ of Σ^* is defined by:

$$\widetilde{\Sigma}_{jk}^{(n)} = \underset{|\sigma|\leq 1}{\operatorname{argmax}}\ L_{jk}^{(n)}(\sigma, \mathbf{y}). \tag{1}$$

Step 2: precision matrix Θ^*

Idea : estimate Θ^* thanks to the graphical Lasso (Friedman et al. (2007)) for the GGM

$\rightsquigarrow$ is to maximize the penalized log-likelihood of the Gaussian model.

$$\hat{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \ \log\det(\Theta) - \operatorname{trace}(\Theta\hat{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\text{off}}. \tag{2}$$

with $\hat{\Sigma}^{(n)}$ being the empirical covariance matrix of $\mathbf{X}$

Step 2: precision matrix Θ^*

Idea : estimate Θ^* thanks to the graphical Lasso (Friedman et al. (2007)) for the GGM

$\rightsquigarrow$ is to maximize the penalized log-likelihood of the Gaussian model.

$$\hat{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \; \log\det(\Theta) - \operatorname{trace}(\Theta\hat{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\text{off}}. \qquad (2)$$

with $\hat{\Sigma}^{(n)}$ being the empirical covariance matrix of $\mathbf{X}$ Problem:

$\Sigma^{(n)}$ unachievable! $\rightsquigarrow$ replaced by $\widetilde{\Sigma}^{(n)}$

Step 2: precision matrix Θ^*

Idea : estimate Θ^* thanks to the graphical Lasso (Friedman et al. (2007)) for the GGM
$\rightsquigarrow$ is to maximize the penalized log-likelihood of the Gaussian model.

$$\hat{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \ \log\det(\Theta) - \operatorname{trace}(\Theta\hat{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\text{off}}. \quad (2)$$

with $\hat{\Sigma}^{(n)}$ being the empirical covariance matrix of $\mathbf{X}$ Problem:

$\Sigma^{(n)}$ unachievable! $\rightsquigarrow$ replaced by $\widetilde{\Sigma}^{(n)}$

Estimator of the precision matrix

Estimator $\widetilde{\Theta}^{(n)}$ of Θ^* is defined by:

$$\widetilde{\Theta}^{(n)} = \underset{\Theta \succ 0}{\operatorname{argmax}} \ \log\det(\Theta) - \operatorname{trace}(\Theta\widetilde{\Sigma}^{(n)}) - \lambda_n ||\Theta||_{1,\text{off}}. \quad (3)$$

Tools

Step 1: Mei & al (2017), *under suitable assumptions, good properties of the population risk can be carried to the empirical risk, even in an non-convex case*
$\rightsquigarrow$ *control of* $|\widetilde{\Sigma}^{(n)}_{j,k} - \Sigma^*_{j,k}|$

Tools

Step 1: Mei & al (2017),
$\rightsquigarrow$ control of $|\widetilde{\Sigma}^{(n)}_{j,k} - \Sigma^*_{j,k}|$

Step 2: Ravikumar & al (2011), *estimating the concentration matrix under sparsity conditions without specific distributional assumptions, but rather analyze the estimator in terms of the tail behavior of* $\max_{j,k} |\widetilde{\Sigma}^{(n)}_{j,k} - \Sigma^*_{j,k}|$

Hypotheses

(H1) For all $j < k$, $|\Sigma^*_{jk}| \neq 1$.

Hypotheses

(H1) For all $j < k$, $|\Sigma^*_{jk}| \neq 1$.

(H2) Let $j < k$. Let g be a function:
$\sigma \in [-1+\delta, 1-\delta] \mapsto \mathbb{E}\Big(L^{(n)}_{jk}(\sigma, \mathbf{y})\Big)$.

- $(-1+\delta)$ et $(1-\delta)$ are not critical points of g,
- g admits a finite number of critical points,
- all the critical points of g, different from Σ^*_{jk}, are non-degenerated, that is:

$$\sigma \neq \Sigma^*_{jk}, \quad g'(\sigma) = 0 \Rightarrow g''(\sigma) \neq 0.$$

Hypotheses

(H1) For all $j < k$, $|\Sigma^*_{jk}| \neq 1$.

(H2) Let $j < k$. Let g be a function:
$\sigma \in [-1+\delta, 1-\delta] \mapsto \mathbb{E}\Big(L^{(n)}_{jk}(\sigma, \mathbf{y})\Big)$.

- $(-1+\delta)$ et $(1-\delta)$ are not critical points of g,
- g admits a finite number of critical points,
- all the critical points of g, different from Σ^*_{jk}, are non-degenerated, that is:

$$\sigma \neq \Sigma^*_{jk}, \quad g'(\sigma) = 0 \Rightarrow g''(\sigma) \neq 0.$$

(H3) Technical (not written). The underlying intuition is to limit the influence of "non-edged" terms on edge terms.

Intermediate result on the estimated covariance matrix $\widetilde{\Sigma}^{(n)}$

Proposition

Under **(H1)** and, **(H2)**, let $0 < \rho < 1$. There exist known constants B, C and D such that if n satisfies $\frac{n}{\log n} \geq C \log\left(\frac{B}{\rho}\right)$, then the estimated covariance matrix $\widetilde{\Sigma}^{(n)}$ satisfies:

$$\mathbb{P}\left(||\widetilde{\Sigma}^{(n)} - \Sigma^*||_\infty \geq D\sqrt{\frac{\log n}{n} \log\left(\frac{B}{\rho}\right)} \right) \leq \frac{p(p-1)}{2}\rho,$$

where $||A||_\infty = \max_{j,k \in \{1,\ldots,p\}} |A_{jk}|$ is the infinite norm of matrix $A \in \mathbb{R}^{p^2}$.

$\rightsquigarrow$ Mei *et al.* (2017)

Final result about $\widetilde{\Theta}^{(n)}$

Theorem

Under **(H1)**, **(H2)** et **(H3)**. Let $c > 2$. There exist known constants B, C and D such that for all n satisfying $\frac{n}{\log n} > f(B, C, D, c, \Sigma^*)$ and $\lambda_n \propto \sqrt{\frac{\log n}{n} \log\left(Bp^c\right)}$, we have with probability $1 - \frac{1}{p^{c-2}}$:

(a) $||\widetilde{\Theta}^{(n)} - \Theta^*||_\infty \leq D\sqrt{\frac{\log n}{n} \log\left(Bp^c\right)}$.

(b) $E(\widetilde{\Theta}^{(n)}) \subset E(\Theta^*)$ and vertices (j,k) is correctly detected as soon as: $\left|\Theta^*_{jk}\right| > D\sqrt{\frac{\log n}{n} \log\left(Bp^c\right)}$.

$\rightsquigarrow$ Ravikumar *et al.* (2011)

Simulations design

- $p = 100$ covariates, $n = 500$ observations

Simulations design

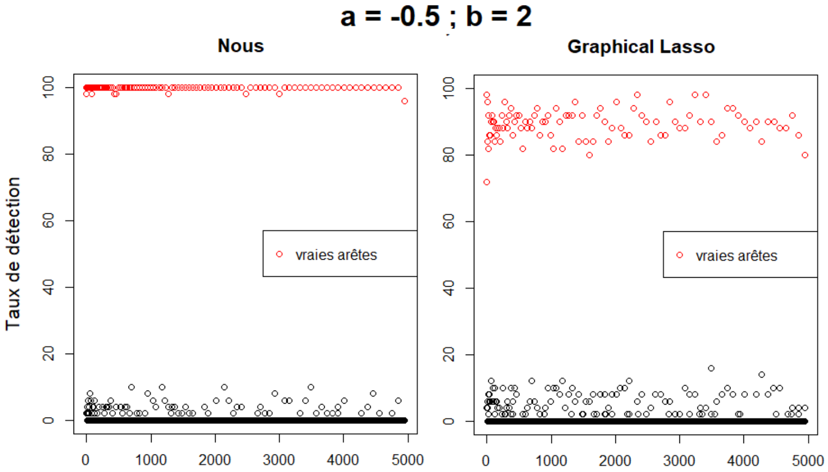
- $p = 100$ covariates, $n = 500$ observations
- Gaussian $\mathbf{X}$ with a chain structure (99 edges) :
  $X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_{100}$

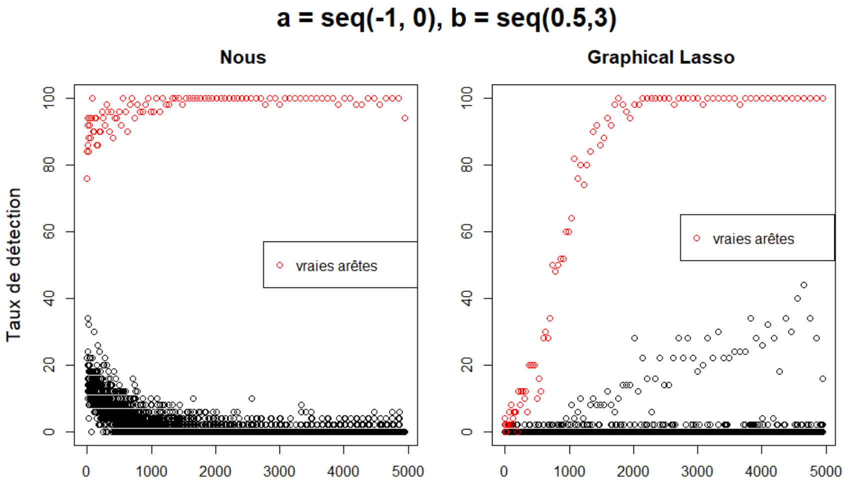
Simulations design

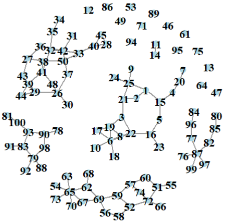
- $p = 100$ covariates, $n = 500$ observations
- Gaussian $\mathbf{X}$ with a chain structure (99 edges) :
  $X_1 \longleftrightarrow X_2 \longleftrightarrow \dots \longleftrightarrow X_{100}$
- Different truncation thresholds:
  - Identical thresholds : $a = -0.5, b = 2$
  - Non identical thresholds $a = seq(-1, 0), b = seq(0.5, 3)$

Simulations design

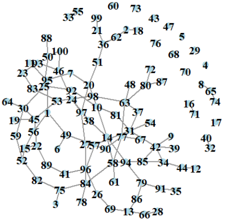
- $p = 100$ covariates, $n = 500$ observations
- Gaussian $\mathbf{X}$ with a chain structure (99 edges) :
  $X_1 \longleftrightarrow X_2 \longleftrightarrow \ldots \longleftrightarrow X_{100}$
- Different truncation thresholds:
  - ➢ Identical thresholds : $a = -0.5, b = 2$
  - ➢ Non identical thresholds $a = seq(-1, 0), b = seq(0.5, 3)$
- Detection rate over 50 repetitions and comparaison with naive graphical Lasso directly on truncated data $\mathbf{Y}$.


a = -0.5 ; b = 2
Nous
Graphical Lasso
Taux de détection
vraies arêtes
vraies arêtes
0
20
40
60
80
100
0
1000
2000
3000
4000
5000

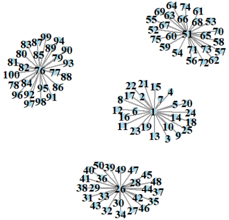

a = seq(-1, 0), b = seq(0.5,3)
Nous
Graphical Lasso
Taux de détection
vraies arêtes
vraies arêtes
0
20
40
60
80
100
0
1000
2000
3000
4000
5000



(a) Structure “cluster”.



(b) Structure “random”.



(c) Structure “hub”.

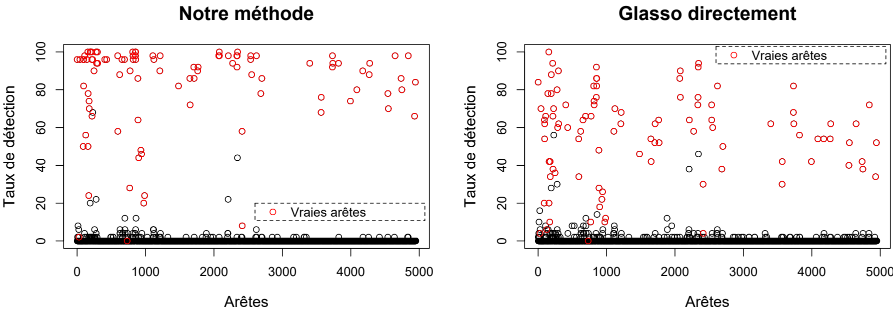


Figure: Detection rates obtained by our method (estimated truncation points) and by the graphical Lasso applied directly to the truncated data. The configurations of the truncation points are "identical only". Detection rates are obtained over 50 independent repetitions for $n = 500$ observations of $p = 100$ variables. "cluster" structure; red = true edges.

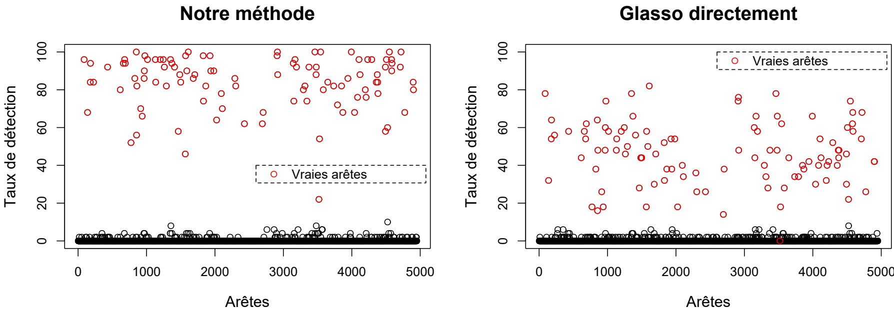


Figure: Detection rates obtained by our method (estimated truncation points) and by the graphical Lasso applied directly to the truncated data. The configurations of the truncation points are "identical only". Detection rates are obtained over 50 independent repetitions for $n = 500$ observations of $p = 100$ variables. "random" structure; red = true edges.

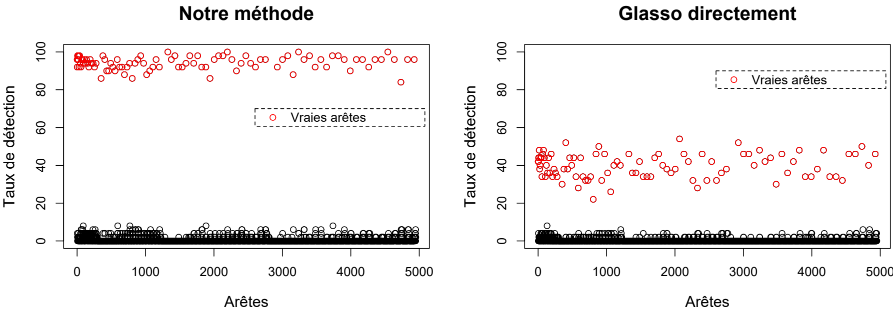


Figure: Detection rates obtained by our method (estimated truncation points) and by the graphical Lasso applied directly to the truncated data. The configurations of the truncation points are "identical only". Detection rates are obtained over 50 independent repetitions for $n = 500$ observations of $p = 100$ variables. "hub" structure; red = true edges.

Grazie mille !

- Mei S., Bei Y., and Montanari A. The landscape of empirical risk for non-convex losses. *arXiv preprint arXiv:1607.06534*, 2017.
- Ravikumar P., Wainwright M. J., Raskutti G., and Yu B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935-980, 2011.